Supplementary Information

Supplementary Note 1: Comparisons to standard software

All timings were performed on a single machine with 128 GB RAM and two Intel Xeon E5-2670 processors, each with 16 threads. All software was run with the maximum number of threads allowed (up to 32) on the 50X Illumina Platinum sequence of NA12878. A standard pipeline including BWA-MEM 0.7.8, Samtools 0.1.18, and Picard Tools 1.99 required 29.2 hours to produce an analysis-ready BAM file. SpeedSeq performs these steps in 7.7 hours, a 3.8-fold speed improvement. The GATK (version 3.2-2-gec30cee) best practices workflow took a total of 36.4 hours for Unified Genotyper and 40.2 hours for Haplotype Caller, even with manual parallelization of Haplotype Caller by chromosome[1]. For SV detection, SpeedSeq runs LUMPY, which is substantially faster than other leading tools. LUMPY took 0.5 hours to call SVs on NA12878, compared with 23.9 hours with Delly v0.5.9 and 18.3 hours with GASVPro (release 2013-10-01)[2,3].

Our implementations of FreeBayes and CNVnator are also faster than their original counterparts. SpeedSeq's parallelized FreeBayes produced SNV and indel calls in 1.4 hours, compared to 19.9 hours for the single-threaded version and 2.2 hours with naïve parallelization by chromosome. CNVnator took only 0.5 hours, compared with 13.5 hours using the original (version 0.3), a 27-fold speed improvement.

Our parallelized software implementations demonstrate little or no difference from the original versions. SAMBLASTER produces nearly identical output to Picard Tools duplicate marking[4]. Our parallelized FreeBayes output differs from the single-threaded standard version at 1,310/4,333,589 sites (0.03%), with only 43 of these variants validating against the GIAB truth set. In our implementation of CNVnator, there were minor differences in output due to the use of a lookup table and a slight loss of floating-point precision. Copy number variable regions differed by ~2,300 bp over the entire genome and often corresponded to ambiguous CNV calls with low-amplitude copy number deviations and low confidence scores.

Supplementary Note 2: Characteristics of miscalled variants

Non-unique 100-mers harbored higher percentages of false positive and false negative SNVs, and both SNVs and indels showed increased false positive rates in satellite repeats (**Supplementary Table 1**). We were not able to evaluate the role of segmental duplications and simple repeats in variant calling accuracy because the Genome in a Bottle truth set excluded these features, however, we expect performance to be similarly diminished in these regions.

Clusters of nearby variants can complicate variant detection due to misalignment, and indeed false positives were enriched in their proximity to indels, with 16.9% of false positive SNVs located within 10 bp of an indel compared to 0.06% of true positives. Upon manual inspection, many of these resulted from complex variants where flanking indels caused SpeedSeq to call phantom SNVs. A smaller portion of miscalls was due to different representations of the same variant, despite our adherence to the variant normalization protocols used by the Genome in a Bottle Consortium[5]. This reflects a persistent deficiency in tools to compare VCF files, and suggests that our metrics likely underestimate SpeedSeq's true performance.

To characterize the nature of miscalled SVs, we used their composite long-read/1KGP validation status to assess whether they were true (TP) or false positives (FP), and counted the number of SVs in which either of the breakends overlapped with annotated genomic features. We expressed this count as a percentage of the number of true positives (N=3,388) or false positives (N=3,088) (**Supplementary Table 1**). False positives overlapped more frequently with simple repeats (TP: 10.6%, FP: 29.4%) and segmental duplications (TP: 4.5%, FP: 23.7%). However, we note the important caveat that SVs in repetitive regions of the genome are also less likely to validate by PacBio or Illumina Moleculo long-reads, and may have been under-ascertained by 1KGP SV mapping efforts, and

therefore some of this enrichment may be due to technical deficiencies in validation (**Supplementary Table 1**).

Supplementary Note 3: De novo variant detection

Identification of *de novo* variants from a family trio is a common WGS application and a natural extension of SpeedSeq's multi-sample variant detection functionality. To evaluate the performance of our tools, we defined a "trio" from the CEPH 1463 pedigree in which NA12878 represented the child, NA12878's biological mother (NA12892) represented one parent, and NA12877 (an unrelated individual) represented her father (**Supplementary Fig. 4**). Thus, the truth set of 288,409 *de novo* SNVs were defined by those variants that were present in the NA12878 GIAB 2.17 variant truth set but absent from NA12877 and NA12892, as determined by an independent variant set from Real Time Genomics[6]. Using the default minimum variant quality score of 1 and no additional filtering, SpeedSeq detected 98.3% of the "*de novo* mutations" with a false discovery rate of 5.9%. These putative mutation calls can be further refined within GEMINI (or using ad hoc scripts) by annotating known polymorphisms (e.g., dbSNP, 1KGP, NHLBI-ESP, ExAC), and by filtering for known sources of false positives. For example, simply requiring all three samples to have a minimum read-depth of 30 at putative mutation sites – a built-in filtering parameter in GEMINI – reduced FDR to 3.2%, with a minor effect on sensitivity (97.1%). Functionally relevant *de novo* mutations can be prioritized by filtering on diverse annotations and variant impact scores. Of course, reducing false positives to very low levels – as required for mutation rate studies – is a more difficult problem that requires extensive probabilistic filtering[7,8].

Supplementary Note 4: Excluded regions

Despite the high quality of the human reference genome, artifacts remain in low-complexity regions and unannotated paralogous sequences that delay processing time and confound variant interpretation. Thus, we have excluded a static set of high-depth regions in the human genome from SNV, indel, and SV breakpoint calling modules of SpeedSeq. These regions exhibit an aberrant increase in sequencing coverage depth where reads from disparate parts of the genome accumulate, violating the diploid assumption of downstream variant calling algorithms. Most high-depth regions are caused by mis-assembled regions of the reference genome, where moderate or high copy repeats have been collapsed into a single (or few) copies, causing read pile-ups.

To identify these high-depth regions, we aligned reads from the CEPH 1463 pedigree to GRCh37 with BWA-MEM and measured aggregate coverage depth from all 17 family members plus one replicate. We then excluded the 15.6 Mb (0.6% of the genome) where the depth was greater than 2 × mode coverage + 3 standard deviations (**Supplementary Fig. 4**). While this static set of excluded regions was based on a single large family, the depth cutoff accommodates a 2-fold increase in copy number relative to the reference genome – corresponding to a homozygous duplication (4 copies) in a genomic region that is single copy in the reference – minimizing the bias toward fixed polymorphism in that family. In the future, we envision that excluded genomic regions will be defined by a much larger set of high-quality human genomes from diverse populations. For structural variation detection with LUMPY, we also exclude the mitochondrial genome, which is prone to false positive calls due to extremely high depth. These regions are not excluded from read-depth analysis using CNVnator, so it is possible (albeit technically complicated) to detect CNVs in these regions using SpeedSeq.

We tested for feature enrichment in repetitive elements from RepeatMasker and mappability from the CRG Alignability track[9,10]. As expected, the excluded regions are highly enriched in repetitive regions of the genome that are known to be poorly assembled in the reference genome including satellite repeats (54-fold enrichment), segmental duplications (9-fold), and sequences near centromeres (28-fold) and other assembly gaps (25-fold) (**Supplementary Table 5**).

	SNVs			Indels			SVs	
Feature	TP N=2,798,941	FP N=12,070	FN N=3,974	TP N=327,165	FP N=3,529	FN N=38,527	TP N=3,388	FP N=3,308
Long interspersed nuclear elements (LINE)	22.5%	18.7%	21.3%	23.6%	18.4%	14.5%	23.3%	20.6%
Short interspersed nuclear elements (SINE)	16.6%	19.6%	40.9%	21.4%	21.8%	34.3%	41.8%	24.5%
Long terminal repeat elements (LTR)	10.5%	16.8%	6.4%	7.1%	8.6%	5.5%	9.7%	11.5%
DNA repeat elements (DNA)	3.6%	3.2%	1.1%	3.7%	2.6%	2.1%	2.9%	3.2%
Satellite repeats	0.2%	5.8%	0.1%	0.1%	0.9%	0.1%	0.8%	1.5%
Simple repeats (micro-satellites)	-	-	-	-	-	-	10.6%	29.4%
Segmental duplications	-	-	-	-	-	-	4.5%	23.7%
Within non-unique 100-mer	7.5%	13.5%	43.7%	4.6%	4.8%	7.9%	50.0%	40.4%
Within 1 Mb of centromere	1.1%	1.7%	3.0%	1.0%	1.1%	1.4%	2.8%	5.4%
Within 1 Mb of telomere	0.5%	5.7%	0.3%	0.4%	1.5%	0.3%	1.1%	4.3%
Within 10 kb of assembly gap	0.1%	1.6%	0.3%	0.1%	0.3%	0.1%	0.1%	1.4%

Supplementary Table 1: Annotation enrichment of miscalled variants. Cells denote the percentage of true positives (TP), false positives (FP), and false negatives (FN) in each variant class. Validations of SNVs and indels were performed against the Genome in a Bottle (GIAB) truth set, and SVs against a combination of the 1000 Genomes callset and PacBio/Moleculo long-reads. Values for SNVs and indels within simple repeats and segmental duplications are not shown because these features are excluded from the GIAB truth set, and only diallelic variants were interrogated in this experiment. False negatives for SVs are not shown because the long-read validation strategy does not produce such results. Note that the enrichment of true positive SV calls in SINEs is due to the detection of variable SINE insertions.

Sample	Type	Detected	Known	Sensitivity	COSMIC variants detected	COSMIC variants known	COSMIC variant sensitivity
TCGA-B6-A0I6	Breast	74	79	93.7%	2	2	100.0%
TCGA-A6-6141	Colorectal	485	510	95.1%	14	14	100.0%
TCGA-CA-6718	Colorectal	1,280	1,307	97.9%	44	44	100.0%
TCGA-D5-6540	Colorectal	779	819	95.1%	19	20	95.0%
TCGA-13-0751	Ovarian	30	31	96.8%	3	3	100.0%
Overall		2,648	2,746	96.4%	82	83	98.8%

Supplementary Table 2: Sensitivity in detecting somatic mutations in tumor-normal pairs. We analyzed five tumor-normal pairs to assess SpeedSeq's sensitivity in detecting 2,746 somatic mutations that had been previously reported by The Cancer Genome Atlas (TCGA) through deep exome sequencing and validated by an orthogonal method. Variants within genes in the COSMIC cancer census gene set defined the "cancer variants" subset.

	Type	Correctly genotyped	Percent correctly genotyped	Detected	Detection sensitivity	Informative occurrences	Unique variants
All SVs	All	7,203	95.1%	7,578	90.2%	8,397	1,722
	Deletion	5,768	95.5%	6,042	90.8%	6,651	1,342
	Duplication	860	93.8%	917	89.5%	1,025	217
	Inversion	555	92.7%	599	85.9%	697	152
	Distant	20	100.0%	20	83.3%	24	11
Heterozygous SVs	All	6,845	96.6%	7,083	89.9%	7,883	1,505
	Deletion	5,421	96.1%	5,641	90.4%	6,240	1,173
	Duplication	853	97.9%	871	89.2%	976	193
	Inversion	551	100.0%	551	85.7%	643	128
	Distant	20	100.0%	20	83.3%	24	11
Homozygous SVs	All	358	72.3%	495	96.3%	514	217
	Deletion	347	86.5%	401	97.6%	411	169
	Duplication	7	15.2%	46	93.9%	49	24
	Inversion	4	8.3%	48	88.9%	54	24
	Distant	0	-	0	-	0	0

Supplementary Table 3: Detection sensitivity and genotyping accuracy of structural variants in the CEPH 1463 grandchildren. SNV-based haplotype phasing of genomic segments in the CEPH 1463 produced 1,722 variants (1,505 heterozygous and 217 homozygous) that were predicted in the 11 grandchildren.

	Number of SVs	Number of deletions	Number of deletions validated	Percent of deletions validated
≥ 7 support	8,456	5,540	4,369	78.9%
Monomorphic	2,525	1,356	1,151	84.9%
Polymorphic	5,931	4,184	3,218	76.9%
Mendelian transmission	5,509	3,853	3,047	79.1%
Mendelian violation	422	331	171	51.7%
Traceable from grandparent	1,722	1,342	1,059	78.9%

Supplementary Table 4: Filtering schematic for 8,456 SVs detected by SpeedSeq in the CEPH 1463 pedigree. Here we show the number of SVs and their validation efficiency at each stage of filtering to produce the 1,722 SVs (bottom left) that were used in SV benchmarking (**Supplementary Table 3**).

Feature	% of mappable genome	% of SpeedSeq excluded regions	Fold enrichment
Long interspersed nuclear elements (LINE)	22.2%	12.8%	0.6
Short interspersed nuclear elements (SINE)	13.9%	6.7%	0.5
Long terminal repeat elements (LTR)	9.3%	5.2%	0.6
DNA repeat elements (DNA)	3.5%	1.5%	0.4
Satellite repeats	0.5%	25.3%	54.0
Simple repeats (micro-satellites)	0.9%	2.7%	2.9
Segmental duplications	5.8%	54.3%	9.4
Within non-unique 100mer	10.7%	57.3%	5.4
Within 1 Mb of centromere	1.3%	36.4%	27.8
Within 1 Mb of telomere	1.3%	4.8%	3.6
Within 10 kb of assembly gap	0.3%	6.4%	24.7

Supplementary Table 5: Annotation enrichment of the SpeedSeq excluded regions. Cells denote the percentage of the mappable genome contained in each annotation class and the percentage of base pairs in the 15.6 Mb SpeedSeq excluded regions contained in the annotation track, along with the fold enrichment of the annotation class in the SpeedSeq excluded regions.

References

1. DePristo, M. A. *et al. Nature Genetics* **43,** 491–498 (2011).
2. Rausch, T. *et al. Bioinformatics* **28,** i333–i339 (2012).
3. Sindi, S. S., Onal, S., Peng, L. C., Wu, H.-T. & Raphael, B. J. *Genome Biol* **13,** R22 (2012).
4. Faust, G. G. & Hall, I. M. *Bioinformatics* **30,** 2503–2505 (2014).
5. Zook, J. M. *et al. Nat Biotechnol* **32,** 246–251 (2014).
6. Cleary, J. G. *et al. J. Comput. Biol.* **21,** 405–419 (2014).
7. Kong, A. *et al. Nature* **488,** 471–475 (2012).
8. Ramu, A. *et al. Nat Meth* **10,** 985–987 (2013).
9. Karolchik, D. *et al. Nucleic Acids Research* **42,** D764–70 (2014).
10. Derrien, T. *et al. PLoS ONE* **7,** e30377 (2012).